Unilever Text Analysis

Julien Maudet, Ian Johnson

Achievements so far

Summarization

Keyphrase Extraction

Web Interface

Summarization

Lemmatization —>
—> Bigrams —>
—> Stopwords —>
—> Vectorization —>
—> SVD —>
—> Semantic Volume Maximization (Yogotama et al.)

Todo

- **Synonyms**
- **Tree parse - smart splitting on conjunctions / delimiters**

Keyphrase extraction

RAKE Algorithm

Introduction

Source: Rose, Stuart, et al. "Automatic keyword extraction from individual documents." *Text Mining* (2010): 1-20.

- Unsupervised & independent from the language at use
- More computationally efficient than TextRank
- Higher precision and comparable recall scores

Keyphrase extraction

RAKE Algorithm

First observation

- keywords = multiple words but no punctuation or stop words (and, the…)

Input

- document
- list of stop words and phrase delimiters
- parameters:
  - minimum length of a word in a keyphrase
  - minimum frequency for a word in the text
  - maximum number of words per keyphrase

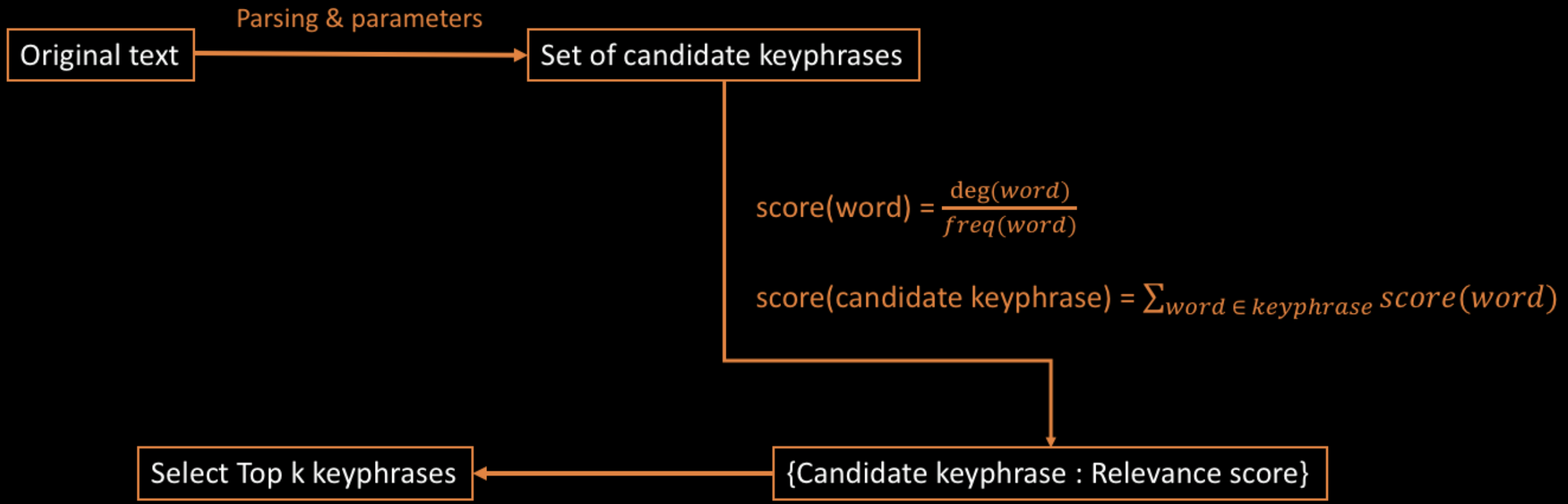
Output

- List of keyphrases and the associated relevance score

Keyphrase extraction

RAKE Algorithm

Pipeline



Web Interface

Technologies

Server side

Python (Algorithms)
Flask framework

User Interface

HTML
Javascript
CSS

Web Interface

LIVE DEMO